

# Nonparametric Bayesian methods for one-dimensional diffusion models: an overview of recent developments

Harry van Zanten  
Korteweg-de Vries Institute for Mathematics  
University of Amsterdam  
hvzanten@uva.nl

October 1, 2012 (this version)

## Abstract

In this paper we review recently developed methods for nonparametric Bayesian inference for one-dimensional diffusion models. We discuss different possible prior distributions, computational issues, and asymptotic results.

## 1 Introduction

Stochastic Langevin models or diffusion models arise in many fields of applied science. Basically they describe the evolution of a system whose dynamics are governed by a “noisy” ordinary differential equation. If  $X_t \in \mathbb{R}^d$  denotes the state of the system at time  $t$ , then the general form of such an equation is

$$\frac{dX_t}{dt} = b(t, X_t) + \text{“Gaussian white noise”}. \quad (1.1)$$

The function  $b$  describes the instantaneous drift of the process  $X$ . This drift is perturbed by random noise, with an intensity that can be time and state dependent in general. A more detailed, but still informal description and many examples of applications in the context of molecular modeling can for instance be found in Section 14.4 of [Schlick \(2010\)](#).

As explained in Section 14.5 of [Schlick \(2010\)](#), we can view (1.1) as an informal description of the stochastic differential equation (SDE)

$$dX_t = b(t, X_t) dt + \sigma(t, X_t) dW_t. \quad (1.2)$$

Here  $W$  is a Brownian motion, which models the random noise, and the diffusion function  $\sigma$  describes the impact of the current time and state on the level of

the noise (see the next section for more details). In this paper we restrict our attention to one-dimensional models, i.e. models for which the state  $X_t$  is real-valued. Moreover, we consider only time-homogenous SDEs, meaning that  $b$  and  $\sigma$  are only functions of the state. Such one-dimensional diffusion models are applied in many different fields in the biosciences. Often they arise after a reduction of a higher-dimensional system to dimension one, achieved for instance by suitable aggregation of data, or by a principle component analysis. Some concrete examples include models for the membrane potential in a neuron (e.g. Lánský et al. (1990)), population size models (Fleming (1975)), decision making models (Roxin and Ledberg (2008)), reduced models for neurodynamical data (Deco et al. (2009)) and models in molecular dynamics (Papaspiliopoulos et al. (2012)).

Fitting a diffusion model to observed data amounts to estimating the functions  $b$  and  $\sigma$ . In certain cases it is reasonable to postulate a specific parametric form of these functions, i.e. to assume that they are known up to some finite-dimensional parameter. A well-known example is the mean-reverting Ornstein-Uhlenbeck process, for which  $\sigma$  is constant and  $b(x) = \alpha(x - \beta)$  for some  $\alpha < 0$  and  $\beta \in \mathbb{R}$ . This is the classical Langevin equation, cf. Langevin (1908). Fitting this model reduces to estimating the parameter  $\theta = (\alpha, \beta, \sigma) \in \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+$ . See for instance Kutoyants (2004) and Kessler et al. (2012) for an overview of methods. In other cases however, natural parametric specifications are not possible or undesirable and one has to resort to nonparametric methods for making inference on the functions  $b$  and  $\sigma$ . Several such methods have been proposed in the literature. An incomplete list include kernel methods (e.g. Banon (1978), Kutoyants (2004), Van Zanten (2001)), penalized likelihood methods (e.g. Comte et al. (2007)), and spectral approaches Bandi and Phillips (2003).

The statistical techniques mentioned thus far are all “frequentist” in nature, i.e. non-Bayesian. In almost all branches of applied statistics however, the use of Bayesian methods has hugely increased in recent years. This is to a large extent due to the development of computational methods. Another appeal is that a Bayes procedure provides a natural way to quantify the uncertainty in the estimates, through the spread of the posterior distribution. Moreover, the construction of a prior distribution, as required in the Bayesian paradigm, can be a useful tool to bring structure to a complex statistical model. Initially the use of Bayes methods in nonparametric problems was met with skepticism, since it was pointed out early on that prior specification is a delicate matter in these cases (e.g. Freedman (1963, 1965)). However, mathematical and practical insights of the last decade have shown that these difficulties can be overcome. As a result Bayesian methods are now widely used in problems like nonparametric regression, density estimation and classification, in many different application areas (see for instance Hjort et al. (2010) for an overview).

The development and study of Bayesian methodology for SDEs started relatively late and initially focussed on parametric models. See for instance the papers Eraker (2001), Roberts and Stramer (2001), Beskos et al. (2006a), to mention but a few. Only a handful of papers about nonparametric Bayes methods for SDEs are available at the present time. The first paper to propose a practical method is Papaspiliopoulos et al. (2012). The theoretical, asymptotic behavior of the procedure of Papaspiliopoulos et al. (2012) is studied in the paper Pokern et al. (2012). Schauer et al. (2012) recently proposed an alternative computational approach. Other available papers deal with asymptotics in this

framework, cf. [Van der Meulen et al. \(2006\)](#), [Panzar and Van Zanten \(2009\)](#) and [Van der Meulen and Van Zanten \(2012\)](#), [Gugushvili and Spreij \(2012\)](#).

In the present paper we review the recently developed nonparametric Bayes methods for SDEs. An interesting aspect of these methods is that they provide a natural way for dealing properly with low-frequency data. Moreover, they allow to report credible bands for uncertainty quantification in addition to an estimator of the function of interest. The few examples that exists at the moment show that the methods have become numerically feasible. As a result, nonparametric Bayes methods can be expected to become more and more common tools for fitting SDE models to observed data.

Throughout the paper a balance is sought between mathematical rigor and a lucid presentation. This means that statements are sometimes loosely formulated, or that regularity conditions are taken for granted. We give references to the literature for readers seeking more mathematical detail.

The remainder of the paper is organized as follows. In the next section we briefly treat some facts from the theory of stochastic differential equations, mainly to recall some terminology and fix notation. In [Section 3](#) we discuss generalities about doing nonparametric Bayesian inference for diffusions. In particular, the differences between continuous and low-frequency data are outlined. Recently proposed concrete priors are considered in [Section 4](#). [Section 5](#) gives an overview of the available asymptotic theory. We end with some concluding remarks in [Section 6](#).

## 2 One-dimensional SDEs

In this section we very briefly review some relevant theory of one-dimensional stochastic differential equations.

### 2.1 Brownian motion

The fundamental building block for stochastic differential equations is the Brownian motion process. Formally, a collection of random variables  $W = (W_t : t \geq 0)$  defined on a common probability space is called a (standard) Brownian motion if

1.  $W_0 = 0$ ,
2. for all  $t \geq s \geq 0$ ,  $W_t - W_s$  is independent of  $(W_u : u \leq s)$ ,
3. for all  $t \geq s \geq 0$ ,  $W_t - W_s$  has a normal distribution with mean 0 and variance  $t - s$ ,
4. almost every sample path  $t \mapsto W_t$  is continuous.

Brownian motion plays a crucial role in stochastic process theory and has been and still is studied extensively. See for instance the books [Revuz and Yor \(1999\)](#) and [Mörters and Peres \(2010\)](#) and the many references therein.

It follows immediately from the definition that the Brownian motion  $W$  is a Gaussian process with mean  $\mathbb{E}W_t = 0$  and covariance  $\mathbb{E}W_s W_t = \min\{s, t\}$  for all  $s, t \geq 0$ . Although it can be proved that the ordinary derivative of Brownian does not exist, it can be thought of as the primitive of Gaussian white noise. As such it is instrumental in putting the loosely described “ODE with noise” [\(1.1\)](#) on a firm mathematical basis.

## 2.2 Stochastic differential equations

Mathematically, the stochastic differential equation (1.2) is shorthand notation for the corresponding integral equation

$$X_t = X_0 + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dW_t. \quad (2.1)$$

Here the second integral is a stochastic integral, which has to be carefully defined and which obeys different calculus rules than ordinary integrals do. In particular, the main rule of calculus is replaced by Itô's formula, which states that if  $X$  solves (1.2) and  $f$  is a twice continuously differentiable function, then  $f(X)$  satisfies the SDE

$$df(X_t) = (f'(X_t)b(t, X_t) + \frac{1}{2}f''(X_t)\sigma^2(t, X_t))dt + f'(X_t)\sigma(t, X_t)dW_t.$$

See any text on stochastic calculus for details (e.g. Chung and Williams (1990), Karatzas and Shreve (1991), Øksendal (2003)). The stochastic integral in (2.1) is well approximated (in probability) by the so-called Riemann-Itô sum

$$\sum_{i=1}^n \sigma((i-1)t/n, X_{(i-1)t/n})(W_{it/n} - W_{(i-1)t/n})$$

if  $n$  is large enough.

It can be proved that under certain regularity conditions on the functions  $b$  and  $\sigma$ , for instance the classical Lipschitz and linear growth conditions, there exists a unique stochastic process  $X$  which solves the integral equation (2.1) (e.g. Karatzas and Shreve (1991)). Moreover, this solution is adapted to the Brownian motion  $W$ , in the sense that for every  $t \geq 0$ ,  $X_t$  only depends on  $(W_s : s \leq t)$ , i.e. it does not “look into the future”.

The notation (1.2) is reasonable since it correctly describes the infinitesimal behavior of the solution  $X$  of (2.1) in the sense that loosely speaking we have for very small  $h > 0$  that

$$X_{t+h} \approx X_t + b(t, X_t)h + \sigma(t, X_t)(W_{t+h} - W_t).$$

By the properties of the Brownian motion and the fact that  $X$  is adapted, we thus have that in distribution,

$$X_{t+h} \approx X_t + b(t, X_t)h + \sqrt{h}\sigma(t, X_t)Z,$$

where  $Z$  is a standard normal random variable independent of  $X_t$ . This gives a basic method for recursively simulating solutions to SDEs, the so-called Euler scheme. See for instance Kloeden and Platen (1992) for (much) more on this topic.

## 2.3 Girsanov's theorem

Below we will restrict our attention to the solution of a stochastic differential equation of the form

$$dX_t = b(X_t)dt + dW_t, \quad X_0 = x_0.$$

For  $T > 0$ , the restriction  $X = (X_t : t \in [0, T])$  of this process to the interval  $[0, T]$  is a random element of the space  $C[0, T]$  of continuous functions on  $[0, T]$ . As such it generates a distribution, or law,  $\mathbb{P}_b^T$  on this function space, defined by  $\mathbb{P}_b^T(B) = \mathbb{P}(X \in B)$  for (Borel) subsets of  $C[0, T]$ .

Girsanov's theorem implies that the distribution  $\mathbb{P}_b^T$  has a density relative to the distribution  $\mathbb{P}_0^T$  of the Brownian motion  $W = (W_t : t \in [0, T])$ . Moreover, we have an expression for the corresponding Radon-Nikodym derivative, or, in statistical terminology, the likelihood. We have

$$\frac{d\mathbb{P}_b^T}{d\mathbb{P}_0^T}(X) = \exp\left(-\frac{1}{2} \int_0^T b^2(X_t) dt + \int_0^T b(X_t) dX_t\right)$$

(see for instance [Liptser and Shiryaev \(2001\)](#)).

### 3 Bayesian inference for SDEs

Suppose we observe a stochastic process  $X$  which solves a one-dimensional SDE of the form

$$dX_t = b(X_t) dt + \sigma(X_t) dW_t,$$

where  $W$  is a Brownian motion and  $b$  and  $\sigma$  are functions satisfying certain regularity conditions ensuring at least that the SDE has a unique (weak) solution for every initial condition. Say we observe the process up till some time  $T > 0$ .

Estimating the diffusion function  $\sigma^2$  is a degenerate statistical problem, at least if the data are recorded continuously, in the sense that its restriction to the range of the data can be recovered without error. We can use the fact that the quadratic variation of the process  $X$  is given by

$$\langle X \rangle_t = \int_0^t \sigma^2(X_s) ds.$$

The left-hand side of this equation is a measurable function of the data, we have for instance that as  $n \rightarrow \infty$ ,

$$\sum_{i=1}^n (X_{it/n} - X_{(i-1)t/n})^2 \rightarrow \langle X \rangle_t$$

in probability, for every  $t \geq 0$  (e.g. [Jacod and Shiryaev \(2003\)](#)). Assuming therefore that  $\sigma$  is known we can instead of the original process  $X$  consider the transformed process

$$\int_0^{X_t} \frac{1}{\sigma(x)} dx, \quad t \geq 0.$$

By Itô's formula this process has unit diffusion and the statistical problem reduces to making inference about its drift function.

In view of these observations we will assume throughout the remainder of the paper that the SDE under consideration has unit diffusion and focus on estimating the drift. In the case of low-frequency data, the transformation outlined above can not be carried out however, and the extension of the methods we discuss ahead to the case of an unknown diffusion function is not always straightforward. We refer to [Roberts and Stramer \(2001\)](#) for an approach that allows a parametric description of the diffusion function.

### 3.1 Continuous-time observations

Suppose that for  $T > 0$ , we observe the unique solution  $X = (X_t : t \in [0, T])$  of the SDE

$$dX_t = b(X_t) dt + dW_t, \quad X_0 = x_0, \quad (3.1)$$

where  $W$  is a Brownian motion and  $b : \mathbb{R} \rightarrow \mathbb{R}$  is an unknown, continuous drift function. By Girsanov's theorem, the law that the process  $X$  generates on the space  $C[0, T]$  of continuous functions on  $[0, T]$  is equivalent to the Wiener measure on the space (with the appropriate initial condition) and the corresponding likelihood satisfies

$$p(X | b) = \exp\left(-\frac{1}{2} \int_0^T b^2(X_t) dt + \int_0^T b(X_t) dX_t\right) \quad (3.2)$$

(see Section 2.3).

The nonparametric Bayesian approach now consists in putting a prior distribution  $\Pi$  on the “parameter”  $b$  and computing the corresponding posterior distribution. Formally the prior  $\Pi$  can be any probability measure on the space  $C(\mathbb{R})$  of continuous functions on  $\mathbb{R}$  or on a suitable subspace, for instance a space of functions with a certain regularity, and/or certain periodicities. The posterior distribution of  $b$ , which we denote by  $\Pi(\cdot | X)$ , is then given by the usual Bayes formula

$$\Pi(b \in B | X) = \frac{\int_B p(b | X) \Pi(db)}{\int p(b | X) \Pi(db)}.$$

(In concrete situations it has to be verified that the integrands are properly measurable, so that the integrals are well defined.)

The integrals in the expression for the posterior are over infinite-dimensional spaces, which makes it challenging to do computations. We will see in Section 4 ahead however that various sensible choices for the prior allow the construction of feasible algorithms for drawing realizations from the posterior.

In general the drift  $b$  in (3.1) is a function defined on the whole real line. This makes it not completely obvious to come up with reasonable priors, as most priors available in the literature are defined on spaces of compactly supported functions. However, we can usually work with such more common priors in the SDE case as well. In certain applications it is natural to assume that  $b$  is a periodic function, reducing the problem to estimating a function on  $[0, 2\pi]$ , or another finite interval. This is for instance the case if the available data consist of recordings of angles (e.g. Pokern (2007), Hindriks (2011) or Papaspiliopoulos et al. (2012)). But also if periodicity can not be assumed it is typically only sensible to estimate the function  $b$  on a compact interval  $I \subset \mathbb{R}$ , since far away from the range of the data there is simply no information available about the function. Now note that if we define, for a set  $S \subset \mathbb{R}$ ,

$$b_S(x) = \begin{cases} b(x) & \text{if } x \in S, \\ 0 & \text{else,} \end{cases}$$

then, with  $I^c$  denoting the complement of the interval  $I$ , the likelihood factorizes as  $p(X | b) = p(X | b_I) p(X | b_{I^c})$ . It follows that if we put a prior on  $b$  by putting

independent priors  $\Pi_I$  and  $\Pi_{I^c}$  on  $b_I$  and  $b_{I^c}$ , respectively, then the marginal posterior for  $b_I$  does not depend on the prior  $\Pi_{I^c}$  and is given by

$$\Pi(b_I \in B \mid X) = \frac{\int_B p(b_I \mid X) \Pi_I(db)}{\int p(b_I \mid X) \Pi_I(db)}.$$

Hence in this case as well, we only need to specify a prior on a compactly supported function.

In the examples in Section 4 we shall consider the periodic case, but simple modifications allow to deal with the non-periodic case as well.

### 3.2 Low-frequency observations

In the preceding subsection we have been dealing with continuously observed diffusions. Obviously, the phrase “continuous data” should be interpreted properly. In practice it means that the frequency at which the diffusion is observed is so high that the error that is incurred by approximating the stochastic and ordinary integrals like the ones appearing in the likelihood (3.2) by the corresponding Riemann or Riemann-Itô sums, is negligible. If we only have low-frequency, discrete-time observations at our disposal, these approximation errors can typically not be ignored however and can introduce undesired biases.

Assume now that we only have partial observations  $X_0, X_\Delta, \dots, X_{n\Delta}$  of the solution of (3.1), for some  $\Delta > 0$  and  $n \in \mathbb{N}$ . We set  $T = n\Delta$ . Under mild regularity conditions the discrete observations constitute a Markov chain, but it is well known that the transition densities of discretely observed diffusions and hence the likelihood is not available in closed form in general. This complicates a Bayesian analysis. An approach that has been proven to be very fruitful is to view the continuous diffusion segments between the observations as missing data and to treat them as latent (function-valued) variables. Since the continuous-data likelihood is known (cf. the preceding subsection), data augmentation methods (see [Tanner and Wong \(1987\)](#)) can be used to circumvent the unavailability of the likelihood in this manner.

Concretely, let us again denote the full, continuous-time process up till time  $T$  by  $X = (X_t : t \in [0, T])$ . Assume that we have solved the inference problem for the continuous data problem described in the preceding subsection, in the sense that we have an algorithm that generates (approximate) draws from the posterior distribution  $\Pi(\cdot \mid b)$  of  $b$  given the full data  $X$ . The data augmentation method relies on the fact that in the present situation it is also possible to generate draws from the conditional distribution

$$X \mid b, X_0, X_\Delta, \dots, X_{n\Delta}$$

of the full process  $X$  given the discrete-time observations  $X_0, X_\Delta, \dots, X_{n\Delta}$  (details ahead). Approximate draws from the target posterior distribution, i.e. the distribution of the drift given  $X_0, X_\Delta, \dots, X_{n\Delta}$ , can then be obtained from a Gibbs sampler which is initialized at some function  $b$  and then repeats the steps

1. draw  $X \mid b, X_0, X_\Delta, \dots, X_{n\Delta}$ ,
2. draw  $b \mid X$

a large number of times.

By the Markov property of the diffusion, step 1. in the Gibbs sampler can be done by independently drawing the  $n$  missing segments

$$(X_t : t \in ((i-1)\Delta, i\Delta)) | b, X_{(i-1)\Delta}, X_{i\Delta} \quad (3.3)$$

for  $i = 1, \dots, n$ , and gluing them together to obtain the full path  $X$ . The crucial observation is that the diffusion bridge law (3.3) is equivalent to a Brownian bridge that starts in  $X_{(i-1)\Delta}$  at time  $(i-1)\Delta$  and ends up in  $X_{i\Delta}$  at time  $i\Delta$ . Moreover, by Girsanov's theorem again, the corresponding Radon-Nikodym derivative is proportional to

$$\exp \left( \int_{(i-1)\Delta}^{i\Delta} b(X_t) dX_t - \frac{1}{2} \int_{(i-1)\Delta}^{i\Delta} b^2(X_t) dt \right).$$

Since it is straightforward to simulate Brownian bridges, this makes it possible to simulate diffusion bridges using for instance rejection sampling or Metropolis-Hastings techniques. Exact simulation methods for diffusion bridges have been proposed in the literature as well, see for instance Beskos et al. (2006a), Beskos et al. (2006b). For the present purposes exact simulation is not strictly necessary however and it is usually more convenient to add a Metropolis-Hastings (MH) step corresponding to a Markov chain that has the diffusion bridge law given by (3.3) as stationary distribution. For more details on this type of MH samplers for diffusion bridges we refer to Roberts and Stramer (2001).

## 4 Gaussian and conditionally Gaussian priors

For successful Bayesian inference for SDEs it is obviously important that a prior is used that makes the procedure computationally feasible. Moreover, to avoid inconsistency problems, we should aim at using priors with “large support”, in the sense that they do not exclude too many drift functions. In this section we review recently proposed options.

### 4.1 Finite-dimensional priors

A first, perhaps naive approach to constructing a prior on the drift function  $b$  is to choose a finite set of basis functions  $\psi_1, \dots, \psi_m$ , assume that the drift admits an expansion  $b = \sum_{j=1}^m c_j \psi_j$  and to put a prior distribution on the vector of coefficients  $c = (c_1, \dots, c_m)$ . In terms of  $c$  the likelihood can then be written as

$$p\left(\sum c_j \psi_j \mid X\right) = e^{c^T \mu - \frac{1}{2} c^T \Sigma c},$$

where the data enters through the vector  $\mu$  and matrix  $\Sigma$  with components

$$\mu_j = \int_0^T \psi_j(X_t) dX_t, \quad \Sigma_{ij} = \int_0^T \psi_i(X_t) \psi_j(X_t) dt,$$

for  $i, j = 1, \dots, m$ . Hence if the prior on the coefficients  $c$  has a Lebesgue density  $\pi$ , then the posterior distribution of  $c$  has a density proportional to

$$c \mapsto \pi(c) e^{c^T \mu - \frac{1}{2} c^T \Sigma c}.$$



Given draws from this posterior for  $c$  we obtain draws from the posterior for  $b$  by combining the coefficients with the basis functions.

Since the likelihood is log-quadratic, it is convenient to choose a Gaussian prior on  $c$ . It is straightforward to verify that if the prior on  $c$  is  $N_m(0, \Lambda)$ , i.e. an  $m$ -dimensional normal distribution with mean 0 and covariance matrix  $\Lambda$ , then the posterior for  $c$  is Gaussian as well, namely  $N_m((\Sigma + \Lambda^{-1})^{-1}\mu, (\Sigma + \Lambda^{-1})^{-1})$ . Sampling from this posterior distribution is straightforward in principle, although the necessary matrix inversions can become numerically demanding as  $m$  gets large. It can be advantageous to employ basis functions leading to a sparse matrix  $\Sigma$ , in order to speed up the matrix computations.

The sketched procedure can work quite well, but only if the drift is in actual fact (close to) a linear combination of the chosen basis functions. When using a prior on a space of functions with a fixed, finite dimension, only the projection of the true drift on this space can be recovered. This can look quite different than the actual drift. Figure 1 illustrates this point. Here we simulated data from the SDE (3.1) with drift function  $b(x) = -(1/2)x(x-1)(x+1)$ . We defined a prior on  $b$  by dividing the interval  $[-2, 2]$  into 20 subintervals of equal length and writing  $b$  as a linear combination of indicator functions of these intervals, with independent Gaussian coefficients. The lower left-hand panel of Figure 1 visualizes the corresponding posterior. The solid blue line is the posterior mean and the dashed lines describe .95 point wise credible intervals. Clearly, this posterior only gives a very crude picture of the true drift (which is the solid black curve). We note that the credible bands are very wide near  $-2$  and  $2$ , since only very limited data fall into that region, cf. the histogram of the data in the lower right-hand panel of the figure.

In the next two subsections we describe two recently proposed procedures that avoid this problem by employing truly infinite-dimensional prior distributions for the drift, both with large support.

## 4.2 Gaussian priors with differential precision operators

Let us assume that the drift function  $b$  in (3.1) is continuously differentiable, 1-periodic and zero-mean, in the sense that  $\int_0^1 b(x) dx = 0$ . For this situation Papaspiliopoulos et al. (2012) propose a centered Gaussian prior on the space  $L^2[0, 1]$  of square integrable functions on the unit interval. In general, a centered Gaussian measure  $\Pi$  on  $L^2[0, 1]$  is determined by its covariance operator  $\Lambda : L^2[0, 1] \rightarrow L^2[0, 1]$  which has the property that

$$\int \left( \int_0^1 g(x) f(x) dx \right) \left( \int_0^1 h(x) f(x) dx \right) \Pi(df) = \int_0^1 g(x) (\Lambda f)(x) dx$$

for all  $g, h \in L^2[0, 1]$ . A linear operator on  $L^2[0, 1]$  is a covariance operator of a Gaussian measure if and only if it is positive definite, symmetric, and trace-class (e.g. Bogachev (1998)). The covariance operator of Papaspiliopoulos et al. (2012) is defined through its inverse, the so-called precision operator. Given fixed hyperparameters  $\eta, \kappa > 0$  and  $p \in \{2, 3, \dots\}$  this inverse is the densely defined operator

$$\Lambda^{-1} = \eta ((-\Delta)^p + \kappa I), \quad (4.1)$$

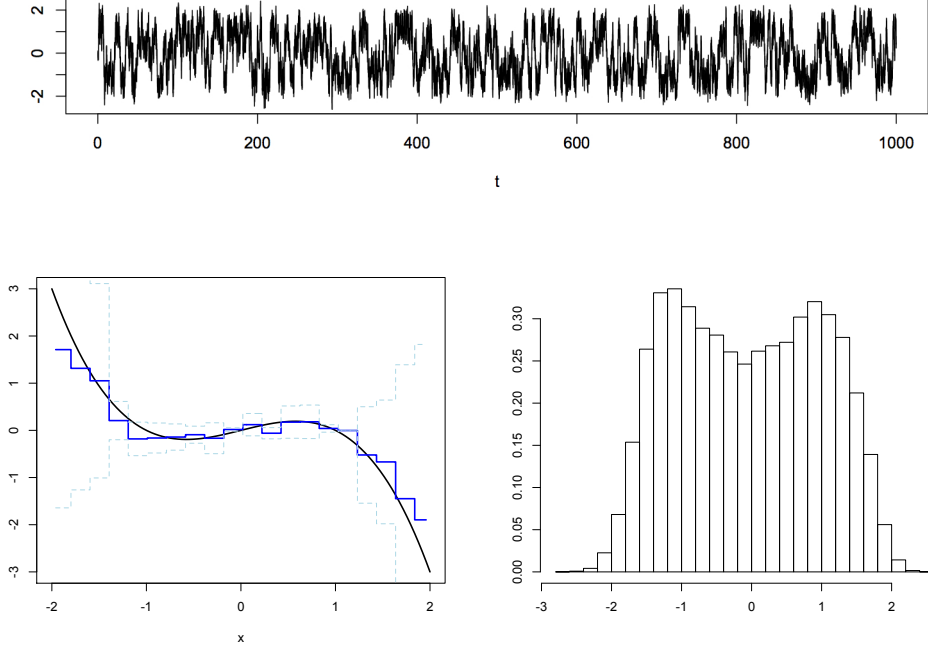


Figure 1: Top panel: simulated data. Lower left: true drift (black), posterior mean (solid blue) and .95 pointwise credible intervals (dashed blue). Lower right: histogram of the data.

where  $\Delta$  the one-dimensional Laplacian, i.e.  $\Delta f = f''$  and  $I$  is the identity operator. The domain of  $\Lambda^{-1}$  is the space of periodic, zero-mean functions with Sobolev regularity  $2p$ . It can be shown that this indeed defines a proper Gaussian prior  $\Pi$  on  $L^2[0, 1]$ . The hyper parameter  $p$  determines the regularity of the prior in some sense. As shown in [Pokern et al. \(2012\)](#), the prior gives mass 1 to a space of functions that have Hölder regularity  $\alpha$  for every  $\alpha < p - 1/2$ .

It turns out that for this prior we can explicitly derive the corresponding posterior. It is a Gaussian measure on  $L^2[0, 1]$  again and we can obtain expressions for posterior mean and precision operator. These expressions involve the periodic local time of the diffusion. This is the random field  $L^\circ = (L_T^\circ(x) : T \geq 0, x \in [0, 1])$  with the property that for every 1-periodic, nonnegative measurable function  $f$  and  $T > 0$ ,

$$\int_0^T f(X_t) dt = \int_0^1 f(x) L_T^\circ(x) dx.$$

Moreover, we need the random field  $\chi^\circ$  defined by

$$\chi_T^\circ(x) = \begin{cases} \#\{k \in \mathbb{Z} : X_0 < x + k < X_T\} & \text{if } X_0 < X_T, \\ -\#\{k \in \mathbb{Z} : X_T < x + k < X_0\} & \text{if } X_T < X_0, \\ 0 & \text{else.} \end{cases}$$

A non-rigorous derivation of the posterior now proceeds by first noting that by Itô's formula and periodicity, the likelihood (3.2) can be written as

$$p(X | b) = \exp \left( -\frac{1}{2} \int_0^1 (L_T^\circ(x) (b^2(x) + b'(x)) - 2\chi_T^\circ(x)b(x)) dx \right).$$

Next we observe that, very loosely speaking, the Gaussian prior  $\Pi$  has a “density” proportional to

$$b \mapsto \exp \left( -\frac{1}{2} \int_0^1 b(x) \Lambda^{-1} b(x) dx \right).$$

But then the posterior has a “density” that is proportional to the product of these two quantities. Using also integration by parts we see that this equals

$$\exp \left( -\frac{1}{2} \int_0^1 b(x) (\Lambda^{-1} + L_T^\circ(x)) b(x) dx + \int_0^1 b(x) \left( \frac{1}{2} (L_T^\circ(x))' + \chi_T^\circ(x) \right) dx \right).$$

This is, up to a constant, again the “density” of a Gaussian measure. By completing the square we see that its mean  $\hat{b}_T$  and covariance operator  $\Lambda_T$  satisfy  $\Lambda_T^{-1} = \Lambda^{-1} + L_T^\circ I$  and  $\Lambda_T^{-1} \hat{b}_T = \frac{1}{2} (L_T^\circ)' + \chi_T^\circ$ . Recalling the definition of  $\Lambda$  we obtain the relations

$$\begin{aligned} \Lambda_T^{-1} &= \eta(-\Delta)^p + (\eta\delta + L_T^\circ) I, \\ \eta(-\Delta)^p \hat{b}_T + (\eta\delta + L_T^\circ) \hat{b}_T &= \frac{1}{2} (L_T^\circ)' + \chi_T^\circ. \end{aligned}$$

To make the derivation of the posterior rigorous the above differential equation for the posterior mean has to be understood in an appropriate weak sense, since the ordinary derivative of local time does not exist. As detailed in [Pokern et al. \(2012\)](#) this is indeed possible and it can be shown that the differential equation for  $\hat{b}_T$  has a unique weak solution. Well-established methods from numerical analysis can be then used to compute the posterior. For further details and application of the approach in problems from molecular dynamics and finance, we refer to [Papaspiliopoulos et al. \(2012\)](#).

The approach of [Papaspiliopoulos et al. \(2012\)](#) can be compared to the naive approach outlined in the preceding subsection by noting that their prior  $\Pi$  can in fact equivalently be described by a series expansion. Define the basisfunctions

$$\begin{aligned} \psi_{2k}(x) &= \sqrt{2} \cos(2\pi kx), \\ \psi_{2k-1}(x) &= \sqrt{2} \sin(2\pi kx), \end{aligned}$$

for  $k \in \mathbb{N}$ . It is easily verified that the functions  $\psi_k$  are the eigenfunctions of the prior precision operator  $\Lambda^{-1}$ , and the corresponding eigenvalues are given by

$$\lambda_k^{-1} = \eta \left( 4\pi^2 \left\lceil \frac{k}{2} \right\rceil^2 \right)^p + \delta.$$

It follows that the prior on  $b$  can also be defined structurally by expanding  $b = \sum_{k=1}^{\infty} c_k \psi_k$  and putting independent Gaussian priors on the coefficients  $c_k$ , with mean 0 and variance  $\lambda_k$ .

So compared to the naive finite series approach, the proposal of Paspiliopoulos et al. (2012) is genuinely nonparametric and eliminates the chance of misspecifying the form of the drift. On the down side, the prior has fixed hyper parameters that still might combine poorly with the true drift. In particular, there remains a possibility that a multiplicative scaling parameter  $\eta$  is chosen that incorrectly reflects the scale of the actual drift. This can deteriorate the quality of the inference. In the next subsection we discuss a prior which allows for a data-driven choice of the scaling.

### 4.3 Infinite series priors

In this section we consider the approach proposed by Schauer et al. (2012). We consider the same setup as before, i.e. the drift  $b$  is assumed to be 1-periodic.

The prior of Schauer et al. (2012) is defined by writing a truncated series expansion for  $b$  and putting prior weights on the truncation point and on the coefficients in the expansion. We employ general 1-periodic, continuous basis functions  $\psi_k$ ,  $k \in \mathbb{N}$ . Next we fix an increasing sequence of natural numbers  $m_j$ ,  $j \in \mathbb{N}$ , to group the basis functions into *levels*. The functions  $\psi_1, \dots, \psi_{m_1}$  constitute level 1, the functions  $\psi_{m_1+1}, \dots, \psi_{m_2}$  correspond to level 2, etcetera. In this manner we can accommodate both families of basis functions with a single index (e.g. the Fourier basis) and doubly indexed families (e.g. wavelet-type bases). Functions that are linear combinations of the first  $m_j$  basis functions  $\psi_1, \dots, \psi_{m_j}$  are said to belong to *model  $j$* . Model  $j$  encompasses levels 1 up till  $j$ .

To define the prior on  $b$  we first put a prior on the model index  $j$ , given by certain prior weights  $p(j)$ ,  $j \in \mathbb{N}$ . By construction, a function in model  $j$  can be expanded as  $\sum_{l=1}^{m_j} \theta_l^j \psi_l$  for a certain vector of coefficients  $\theta^j \in \mathbb{R}^{m_j}$ . Given  $j$ , we endow this vector with a prior by postulating that the coefficients  $\theta_l^j$  are given by an inverse gamma scaling constant times independent, centered Gaussians with certain decreasing variances.

Concretely, to define the prior we fix model probabilities  $p(j)$ ,  $j \in \mathbb{N}$ , decreasing variances  $\xi_l^2$ ,  $l \in \mathbb{N}$ , positive constants  $a, b > 0$  and set  $\Xi_j = \text{diag}(\xi_1^2, \dots, \xi_{m_j}^2)$ . Then the hierarchical prior  $\Pi$  on the drift function  $b$  is defined as follows:

$$\begin{aligned} j &\sim p(j), \\ s^2 &\sim \text{IG}(a, b), \\ \theta^j \mid j, s^2 &\sim N_{m_j}(0, s^2 \Xi_j), \\ b \mid j, s^2, \theta^j &\sim \sum_{l=1}^{m_j} \theta_l^j \psi_l. \end{aligned}$$

In the paper Schauer et al. (2012) particular choices for the basis functions, the prior on  $j$  and the variances  $\xi_l$  are considered in more detail.

This prior proposed by Schauer et al. (2012) is different from the one considered above in a number of ways. Firstly, different basis functions can be used. This added flexibility can be computationally attractive. The posterior computations involve the inversion of certain large matrices and choosing basis

functions with local support typically makes these matrices sparse. A second difference is that we truncate the infinite series at a level that we endow with a prior as well. In this manner we can achieve considerable computational gains if the data driven truncation point is relatively small, so that only low-dimensional models are used and hence only relatively small matrices have to be inverted. A last important change is that we do not set the multiplicative hyper parameter at a fixed value, but instead endow it with a prior and let the data determine the appropriate value.

The simulations presented in [Schauer et al. \(2012\)](#) indicate that this approach has several advantages. Although the truncation of the series at a data driven point involves incorporating reversible jump MCMC steps in the computational algorithm, it can indeed lead to a considerably faster procedure compared to truncating at some fixed high level. Moreover, the introduction of a prior on the multiplicative hyper parameter reduces the risk of misspecifying the scale of the drift. Using a fixed scaling parameter can seriously deteriorate the quality of the inference, whereas the hierarchical procedure with a prior on that parameter is able to adapt to the true scale of the drift. Also, numerical investigations indicate that by the introduction of a prior on the scale we also can achieve some degree of adaptation to smoothness.

The prior  $\Pi$  described above is constructed in such a way that numerical computation is practically feasible. Within a fixed model  $j$ , a Gibbs sampler for sampling  $\theta^j$  and  $s^2$  can be constructed using standard inverse gamma-normal computations. Reversible jump MCMC can be used to jump between different models. This involves the computation of certain Bayes factors, for which a closed form expression can be derived in this setup. If only low-frequency data are available, the Gibbs and reversible MCMC steps can be combined with data augmentation steps involving the simulation of diffusion bridges, as outlined also in the preceding subsection. For more details about computational issues and simulation examples we refer to [Schauer et al. \(2012\)](#).

## 5 Asymptotics

The negative examples of e.g. [Freedman \(1963, 1965\)](#) or [Diaconis and Freedman \(1986a,b\)](#) show that in Bayesian nonparametrics, even intuitively reasonable priors may lead to inconsistent procedures. More generally, it is by now well known that contrary to the parametric setting, the choice of the prior has a large impact on the performance in infinite-dimensional models. This performance is determined by fine mathematical properties and can not be assessed by simply eyeballing the prior. As a result there is an interest in mathematical results that relate properties of the prior to the quality of the Bayes procedure. Such results can serve as guidelines for the selection or construction of priors.

Mathematical results in this setting typically assume that the data are generated using a true drift function  $b_0$  and study if and how the posterior concentrates around  $b_0$  as more and more data become available. In the continuous-time case in which we observe the diffusion on a time interval  $[0, T]$  this means we let  $T \rightarrow \infty$ . In the low-frequency case it simply means we let the number  $n$  of observations tend to infinity. Posterior consistency is the property that the posterior indeed contracts around  $b_0$ , in the sense that asymptotically, any neighborhood (relative to a suitably defined topology) of  $b_0$  receives posterior

mass 1. This is a property that any reasonable procedure should ideally have. Once posterior consistency has been established, the rate at which the posterior contracts around the true  $b_0$  can be studied. In particular it can be investigated whether a certain prior leads to optimal convergence rates.

For diffusion models, the paper [Van der Meulen et al. \(2006\)](#) was the first to systematically study convergence rates for nonparametric Bayes procedures. General conditions were derived for attaining a certain rate of contraction, in terms of the metric entropy of the support of the prior and the mass that the prior assigns to neighborhoods of the true function. These conditions are the analogues of similar conditions that were initially derived for the setting of i.i.d. density estimation by [Ghosal et al. \(2000\)](#). A concrete prior for ergodic diffusions considered in [Van der Meulen et al. \(2006\)](#) is a Dirichlet process like prior designed to model decreasing drift functions. This prior is shown to attain the optimal convergence rate  $T^{-1/3}$  (up to a logarithmic factor). Certain Gaussian process priors for the drift, essentially multiply integrated Brownian motions, are considered in the paper [Panzar and Van Zanten \(2009\)](#). These priors have a certain fixed degree of regularity. Brownian motion has smoothness of order  $1/2$ , integrated Brownian motion has smoothness level  $3/2$ , etc. For such Gaussian priors the message is that if the regularity of the prior that is used coincides with the regularity of the unknown drift function, then optimal contraction rates are achieved. Specifically, if both the drift and the prior have regularity  $\beta > 0$ , then the posterior contracts around the true drift at the rate  $T^{-\beta/(1+2\beta)}$ , which can be shown to be optimal in a certain sense, cf. e.g. [Kutoyants \(2004\)](#). If the two regularities are not equal however, only sub-optimal speeds are realized in general. This is in line with general findings for Gaussian priors in other settings, see e.g. [Van der Vaart and Van Zanten \(2008\)](#), [Castillo \(2008\)](#).

The concrete Gaussian prior with precision operator (4.1) described in Section 4.2 is analyzed in detail in the paper [Pokern et al. \(2012\)](#). As mentioned above, this prior has regularity  $\beta = p - 1/2$ . It is proved in [Pokern et al. \(2012\)](#) that the corresponding posterior contracts around the true drift at the rate  $T^{-(2p-1)/(4p)}$ , provided the drift has regularity  $p$ . Note that this rate is also the optimal  $T^{-\beta/(1+2\beta)}$ , but the assumption on the drift is that it is  $\beta + 1/2$ -regular. It is expected however that also in the periodic setting of [Pokern et al. \(2012\)](#) this assumption on the drift can be weakened to  $\beta$ -regularity, and that just as in the ergodic setting studied in [Panzar and Van Zanten \(2009\)](#), it holds that a Gaussian prior with fixed regularity is rate-optimal if and only if its smoothness matches the smoothness of the true drift.

Priors that perform optimally across a whole range of regularities for the drift, i.e. so-called rate-adaptive priors have not yet been exhibited for diffusion models. It is expected that the prior of [Schauer et al. \(2012\)](#) described in Section 4.3 allows for a degree of adaptation to smoothness, but this has not yet been established. A combination of the general theory of [Van der Meulen et al. \(2006\)](#) and new local time asymptotics obtained in [Pokern et al. \(2012\)](#) are expected to shed further light on the matter in the near future.

The asymptotic results mentioned thus far all concern continuously observed diffusions, where the accumulation of information is ensured either by ergodicity or by periodicity assumptions. The derivation of usable results for the low-frequency setting is much more involved. The fact that the discrete-time likelihood can not be explicitly expressed in terms of the drift complicates the analysis considerably. At the present time, the only available results deal with posterior

consistency relative to a weak topology, cf. [Van der Meulen and Van Zanten \(2012\)](#) and the extensions in [Gugushvili and Spreij \(2012\)](#). It is a great challenge to obtain consistency results for stronger topologies and rate of contraction results for procedures based on low-frequency data.

## 6 Concluding remarks

Nonparametric Bayesian methodology for stochastic differential equations has started to develop only very recently. At this point in time there exist only a few methods that are computationally feasible. Moreover, the theoretical performance analysis of these methods is still rather immature. Nevertheless, the nonparametric Bayes approach can be expected to become more and more common in the near future, since it combines the advantages of flexible, nonparametric modeling with the possibility of providing uncertainty quantification and the possibility to deal with low-frequency data. This development will be stimulated by ongoing work on computational matters and theoretical foundations.

## References

- Bandi, F. M. and Phillips, P. C. B. (2003). Fully nonparametric estimation of scalar diffusion models. *Econometrica* **71**(1), 241–283.
- Banon, G. (1978). Nonparametric identification for diffusion processes. *SIAM J. Control Optim.* **16**(3), 380–395.
- Beskos, A., Papaspiliopoulos, O. and Roberts, G. O. (2006a). Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli* **12**(6), 1077–1098.
- Beskos, A., Papaspiliopoulos, O., Roberts, G. O. and Fearnhead, P. (2006b). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**(3), 333–382.
- Bogachev, V. I. (1998). *Gaussian measures*. American Mathematical Society, Providence, RI.
- Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2**, 1281–1299.
- Chung, K. L. and Williams, R. J. (1990). *Introduction to stochastic integration*. Birkhäuser Boston Inc., Boston, MA, second edition.
- Comte, F., Genon-Catalot, V. and Rozenholc, Y. (2007). Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli* **13**(2), 514–543.
- Deco, G., Mart, D., Ledberg, A., Reig, R. and Sanchez Vives, M. V. (2009). Effective reduced diffusion-models: A data driven approach to the analysis of neuronal dynamics. *PLoS Comput Biol* **5**(12), e1000587.

- Diaconis, P. and Freedman, D. (1986a). On inconsistent Bayes estimates of location. *Ann. Statist.* **14**(1), 68–87.
- Diaconis, P. and Freedman, D. (1986b). On the consistency of Bayes estimates. *Ann. Statist.* **14**(1), 1–67.
- Eraker, B. (2001). MCMC analysis of diffusion models with application to finance. *J. Bus. Econom. Statist.* **19**(2), 177–191.
- Fleming, W. H. (1975). Diffusion processes in population biology. *Advances in Applied Probability* **7**, pp. 100–105.
- Freedman, D. A. (1963). On the asymptotic behavior of Bayes’ estimates in the discrete case. *Ann. Math. Statist.* **34**, 1386–1403.
- Freedman, D. A. (1965). On the asymptotic behavior of Bayes estimates in the discrete case. II. *Ann. Math. Statist.* **36**, 454–456.
- Ghosal, S., Ghosh, J. K. and Van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28**(2), 500–531.
- Gugushvili, S. and Spreij, P. (2012). Non-parametric Bayesian drift estimation for stochastic differential equations. *ArXiv* **1206.4981**.
- Hindriks, R. (2011). *Empirical dynamics of neuronal rhythms*. PhD thesis, VU University Amsterdam.
- Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G., eds. (2010). *Bayesian nonparametrics*, volume 28 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.
- Jacod, J. and Shiryaev, A. N. (2003). *Limit theorems for stochastic processes*, volume 288. Springer-Verlag, Berlin, second edition.
- Karatzas, I. and Shreve, S. E. (1991). *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition.
- Kessler, M., Lindner, A. and Sørensen, M. (2012). *Statistical Methods for Stochastic Differential Equations*. Monographs on Statistics and Applied Probability. Taylor & Francis.
- Kloeden, P. E. and Platen, E. (1992). *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin.
- Kutoyants, Y. A. (2004). *Statistical inference for ergodic diffusion processes*. Springer Series in Statistics. Springer-Verlag London Ltd., London.
- Langevin, P. (1908). On the theory of Brownian motion. *C. R. Acad. Sci.* **146**, 530–533.
- Lánský, P., Smith, C. E. and Ricciardi, L. M. (1990). One-dimensional stochastic diffusion models of neuronal activity and related first passage time problems. *Trends Biol Cybern* **1**, 153–162.



- Liptser, R. S. and Shiryaev, A. N. (2001). *Statistics of random processes. I*, volume 5 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, expanded edition.
- Mörters, P. and Peres, Y. (2010). *Brownian motion*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.
- Øksendal, B. (2003). *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, sixth edition.
- Panzar, L. and Van Zanten, J. H. (2009). Nonparametric Bayesian inference for ergodic diffusions. *J. Statist. Plann. Inference* **139**(12), 4193–4199.
- Papaspiliopoulos, O., Pokern, Y., Roberts, G. O. and Stuart, A. M. (2012). Nonparametric estimation of diffusions: a differential equations approach. *Biometrika* **99**(3), 511–531.
- Pokern, Y. (2007). *Fitting Stochastic Differential Equations to Molecular Dynamics Data*. PhD thesis, University of Warwick.
- Pokern, Y., Stuart, A. M. and Van Zanten, J. H. (2012). Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDEs. To appear in *Stoch. Proc. Appl.*
- Revuz, D. and Yor, M. (1999). *Continuous martingales and Brownian motion*. Springer-Verlag, Berlin, third edition.
- Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika* **88**(3), 603–621.
- Roxin, A. and Ledberg, A. (2008). Neurobiological models of two-choice decision making can be reduced to a one-dimensional nonlinear diffusion equation. *PLoS Comput Biol* **4**(3), e1000046.
- Schauer, M., Van der Meulen, F. and Van Zanten, J. H. (2012). Reversible jump MCMC for nonparametric drift estimation for diffusion processes. *ArXiv* **1206.4910**.
- Schlick, T. (2010). *Molecular modeling and simulation*. Springer, New York, second edition.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82**(398), 528–550.
- Van der Meulen, F. H., Van der Vaart, A. W. and Van Zanten, J. H. (2006). Convergence rates of posterior distributions for Brownian semimartingale models. *Bernoulli* **12**(5), 863–888.
- Van der Meulen, F. H. and Van Zanten, J. H. (2012). Consistent nonparametric Bayesian inference for discretely observed scalar diffusions. *Bernoulli* (to appear).
- Van der Vaart, A. W. and Van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36**(3), 1435–1463.

Van Zanten, H. (2001). Rates of convergence and asymptotic normality of kernel estimators for ergodic diffusion processes. *J. Nonparametr. Statist.* **13**(6), 833–850.